Model-based co-clustering for ordinal data

Julien Jacques, Christophe Biernacki

Model-Based Co-clustering for Ordinal Data

Julien Jacques [1] & Christophe Biernacki [2]

[1] *Univ Lyon, Lumière Lyon 2, ERIC, Lyon, France, julien.jacques@univ-lyon2.fr*
[2] *University Lille 1 et Inria, Lille, France, christophe.biernacki@math.univ-lille1.fr*

Résumé. Nous présentons dans ce travail un algorithme de coclustering pour données ordinales. Cet algorithme repose sur le modèle des blocs latents utilisant le modèle BOS (Biernacki et Jacques, 2015, Stat. Comput.) pour données ordinales et un algorithme SEM-Gibbs pour l'inférence. Des expériences sur données simulées illustrent l'efficacité de la méthode d'estimation.

Mots-clés. coclustering, données ordinales, algorithme SEM-Gibbs

Abstract. A model-based coclustering algorithm for ordinal data is presented. This algorithm relies on the latent block model using the BOS model (Biernacki and Jacques, 2015, Stat. Comput.) for ordinal data and a SEM-Gibbs algorithm for inference. Numerical experiments on simulated data illustrate the efficiency of the inference strategy.

Keywords. model-based coclustering, ordinal data, SEM-Gibbs algorithm

1 Intoduction

Historically, clustering algorithms are used to explore data and to provide a simplified representation of them with a small number of homogeneous groups of individuals (i.e. clusters). With the big data phenomenon, the number of features becomes itself larger and larger, and traditional clustering methods are no more sufficient to explore such data. Coclustering algorithms have been introduced to provide a solution by gathering into homogeneous groups both the observations and the features. Among numerous algorithms recently developed, model-based approaches [2] have proven their efficiency.

This work focuses on ordinal data, which are categorical data with ordered levels. To the best of our knowledge, the first model-based coclustering algorithm for such data has been proposed by [5], relying on the proportional odds model. In this work, we propose a model-based coclustering algorithm relying on the BOS (*Binary Ordinal Search*) model [1], which has proven its efficiency for modeling ordinal data.

The data set is composed of a matrix of n observations (rows or individuals) of d ordinal variables (columns or features): $\mathbf{x} = (x_{ih})_{1\leq i\leq n, 1\leq h\leq d}$. For simplicity, the levels of x_{ih} will be numbered $\{1, ..., m_h\}$, and all m_h's are assumed to be equal: $m_h = m$ $(1 \leq h \leq d)$.

2 Latent block model for ordinal data

Latent block model The latent block model assumes local independence i.e., the $n \times d$ random variables $\mathbf{x}$ are assumed to be independent once the row partition $\mathbf{v} = (v_{ik})_{1 \leq i \leq n, 1 \leq k \leq K}$ and the column partition $\mathbf{w} = (w_{h\ell})_{1 \leq h \leq d, 1 \leq \ell \leq L}$ are fixed (note that a standard binary partition is used for $\mathbf{v}$ and $\mathbf{w}$):

$$\mathrm{p}(\mathbf{x}; \theta) = \sum_{\mathbf{v} \in V} \sum_{\mathbf{w} \in W} \mathrm{p}(\mathbf{v}; \theta) \mathrm{p}(\mathbf{w}; \theta) \mathrm{p}(\mathbf{x} | \mathbf{v}, \mathbf{w}; \theta) \tag{1}$$

with (below the straightforward range for i, h, k and ℓ are omitted):

- V the set of all possible partitions of rows into K groups, W the set of partitions of the columns into L groups,
- $\mathrm{p}(\mathbf{v}; \theta) = \prod_{ik} \alpha_k^{v_{ik}}$ and $\mathrm{p}(\mathbf{w}; \theta) = \prod_{h\ell} \beta_\ell^{w_{h\ell}}$ where α_k and β_ℓ are the row and column mixing proportions, belonging to $[0, 1]$ and summing to 1,
- $\mathrm{p}(\mathbf{x} | \mathbf{v}, \mathbf{w}; \theta) = \prod_{ihk\ell} \mathrm{p}(x_{ih}; \mu_{k\ell}, \pi_{k\ell})^{v_{ik} w_{h\ell}}$ where $\mathrm{p}(x_{ih}; \mu_{k\ell}, \pi_{k\ell})$ is the probability of x_{ij} according to the BOS model [1] parametrized by $(\pi_{k\ell}, \mu_{k\ell})$ with the so-called precision parameter $\pi_{k\ell} \in [0, 1]$ and position parameter $\mu_{k\ell} \in \{1, ..., m\}$ (detail of $\mathrm{p}(x_{ih}; \mu_{k\ell}, \pi_{k\ell})$ is given below),
- $\theta = (\pi_{k\ell}, \mu_{k\ell}, \alpha_k, \beta_\ell)$ is the whole mixture parameter.

Ordinal model The BOS model [1], built using the assumption that an ordinal variable is the result of a stochastic binary search algorithm in which e_j is the current interval in $\{1, \ldots, n\}$, and y_j the break point in this interval, is defined as follows:

$$\mathrm{p}(x_{ij}; \mu_{k\ell}, \pi_{k\ell}) = \sum_{e_{m-1}, \ldots, e_1} \prod_{j=1}^{m-1} \mathrm{p}(e_{j+1} | e_j; \mu_{k\ell}, \pi_{k\ell}) \mathrm{p}(e_1) \tag{2}$$

where

$$\begin{aligned}
\mathrm{p}(e_{j+1} | e_j; \mu_{k\ell}, \pi_{k\ell}) &= \sum_{y_j \in e_j} \mathrm{p}(e_{j+1} | e_j, y_j; \mu, \pi) \mathrm{p}(y_j | e_j), \\
\mathrm{p}(e_{j+1} | e_j, y_j; \mu_{k\ell}, \pi_{k\ell}) &= \pi_{k\ell} \mathrm{p}(e_{j+1} | y_j, e_j, z_j = 1; \mu_{k\ell}) + (1 - \pi_{k\ell}) \mathrm{p}(e_{j+1} | y_j, e_j, z_j = 0), \\
\mathrm{p}(e_{j+1} | y_j, e_j, z_j = 0) &= \frac{|e_{j+1}|}{|e_j|} \mathbb{I}(e_{j+1} \in \{e_j^-, e_j^=, e_j^+\}), \\
\mathrm{p}(e_{j+1} | y_j, e_j, z_j = 1; \mu_{k\ell}) &= \mathbb{I}(e_{j+1} = \operatorname*{argmin}_{e \in \{e_j^-, e_j^=, e_j^+\}} \delta(e, \mu_{k\ell})) \mathbb{I}(e_{j+1} \in \{e_j^-, e_j^=, e_j^+\}),
\end{aligned}$$

with δ a "distance" between μ and an interval $e = \{b^-, \ldots, b^+\}$: $\delta(e, \mu) = \min(|\mu - b^-|, |\mu - b^+|)$ and also

$$\mathrm{p}(z_j|e_j; \pi_{k\ell}) = \pi \mathbb{I}(z_j = 1) + (1 - \pi_{k\ell})\mathbb{I}(z_j = 0) \quad \text{and} \quad \mathrm{p}(y_j|e_j) = \frac{1}{|e_j|}\mathbb{I}(y_j \in e_j).$$

The density (1) can finally be written

$$\mathrm{p}(\mathbf{x}; \theta) = \sum_{\mathbf{v} \in V} \sum_{\mathbf{w} \in W} \prod_{ik} \alpha_k^{v_{ik}} \prod_{h\ell} \beta_\ell^{w_{h\ell}} \prod_{ihk\ell} \mathrm{p}(x_{ih}; \mu_{k\ell}, \pi_{k\ell})^{v_{ik} w_{h\ell}}. \tag{3}$$

Missing data In the present work, the data $\mathbf{x}$ may be also incomplete. We will notice $\check{\mathbf{x}}$ the set of observed data, $\hat{\mathbf{x}}$ the set of unobserved data and $\mathbf{x}$ the set of both observed and unobserved data. The algorithm we propose below will be able to take into account these missing data and to estimate them.

3 Model inference

The aim is to estimate θ by maximizing the observed log-likelihood

$$\ell(\theta; \check{\mathbf{x}}) = \sum_{\hat{\mathbf{x}}} \ln \mathrm{p}(\mathbf{x}; \theta). \tag{4}$$

For computational reasons, and EM algorithm is not feasible in that coclustering case (see [2]), thus we opt for one of its stochastic version denoted by SEM-Gibbs [4].

3.1 SEM-Gibbs algorithm

The proposed SEM-Gibbs algorithm relies on the EM algorithm used in [1] for the estimation of the BOS model. Starting from an initial value for the parameter $\theta^{(0)}$ and for the missing data $\hat{\mathbf{x}}^{(0)}, \mathbf{w}^{(0)}$, the qth iteration of the SEM-Gibbs algorithm alternates the following SE and M steps ($q \geq 0$).

SE step Execute a small number (at least 1) of successive iterations of the following three steps:

1. generate the row partition $v_{ik}^{(q+1)}|\hat{\mathbf{x}}^{(q)}, \check{\mathbf{x}}, \mathbf{w}^{(q)}$ for all $1 \leq i \leq n, 1 \leq k \leq K$:

$$\mathrm{p}(v_{ik} = 1|\hat{\mathbf{x}}^{(q)}, \check{\mathbf{x}}, \mathbf{w}^{(q)}; \theta^{(q)}) = \frac{\alpha_k^{(q)} f_k(x_{i.}^{(q)}|\mathbf{w}^{(q)}; \theta^{(q)})}{\sum_{k'} \alpha_{k'}^{(q)} f_{k'}(x_{i.}^{(q)}|\mathbf{w}^{(q)}; \theta^{(q)})}$$

where $f_k(x_{i.}^{(q)}|\mathbf{w}^{(q)}; \theta^{(q)}) = \prod_{h\ell} \mathrm{p}(x_{ih}^{(q)}; \mu_{k\ell}^{(q)}, \pi_{k\ell}^{(q)})^{w_{h\ell}^{(q)}}$ and $x_{i.}^{(q)} = \{\hat{\mathbf{x}}^{(q)}, \check{\mathbf{x}}\}_i$ denotes observed data $\{\check{\mathbf{x}}\}_i$ completed by the generation of the unobserved one $\{\hat{\mathbf{x}}^{(q)}\}_i$ at the qth step.

2. symmetrically, generate the column partition $w_{h\ell}^{(q+1)}|\hat{\mathbf{x}}^{(q)}, \check{\mathbf{x}}, \mathbf{v}^{(q+1)}$ for all $1 \leq h \leq d, 1 \leq \ell \leq L$:

$$\mathrm{p}(w_{h\ell} = 1|\hat{\mathbf{x}}^{(q)}, \check{\mathbf{x}}, \mathbf{v}^{(q+1)}; \theta^{(q)}) = \frac{\beta_\ell^{(q)} g_\ell(x_{.h}^{(q)}|\mathbf{v}^{(q+1)}; \theta^{(q)})}{\sum_{\ell'} \beta_{\ell'}^{(q)} g_{\ell'}(x_{.h}^{(q)}|\mathbf{v}^{(q+1)}; \theta^{(q)})}$$

where $g_\ell(x_{.h}^{(q)}|\mathbf{v}^{(q+1)}; \theta^{(q)}) = \prod_{ik} \mathrm{p}(x_{ih}^{(q)}; \mu_{k\ell}^{(q)}, \pi_{k\ell}^{(q)})^{v_{ik}^{(q+1)}}$ and $x_{.h}^{(q)} = \{\hat{\mathbf{x}}^{(q)}, \check{\mathbf{x}}\}_h$

3. generate the missing data $\hat{x}_{ih}^{(q+1)}|\check{\mathbf{x}}, \mathbf{v}^{(q+1)}, \mathbf{w}^{(q+1)}$ following

$$\mathrm{p}(\hat{x}_{ih}|\check{\mathbf{x}}, \mathbf{v}^{(q+1)}, \mathbf{w}^{(q+1)}; \theta^{(q)}) \propto \prod_k \alpha_k^{(q)^{v_{ik}^{(q+1)}}} \prod_\ell \beta_\ell^{(q)^{w_{h\ell}^{(q+1)}}} \prod_{k\ell} \mathrm{p}(\hat{x}_{ih}; \mu_{k\ell}^{(q)}, \pi_{k\ell}^{(q)})^{v_{ik}^{(q+1)} w_{h\ell}^{(q+1)}}.$$

M step Estimate θ, conditionally on $\hat{\mathbf{x}}^{(q+1)}, \mathbf{v}^{(q+1)}, \mathbf{w}^{(q+1)}$ obtained at the SE step (and also conditionally to $\check{\mathbf{x}}$), using the EM algorithm of [1].

Choosing the parameter estimation After a burn in period, the final estimation of the discrete parameter $\mu_{k\ell}$ is the mode of the sample distribution, and the final estimation of the continuous parameters $(\pi_{k\ell}, \alpha_k, \beta_\ell)$ is the mean of the sample distribution. It produces a final estimate $\hat{\theta}$. The missing data estimation $\hat{\mathbf{x}}$ are obtained by the mode of the sample distribution after the burn-in period too.

Choosing the partition The final bi-partition $(\hat{\mathbf{v}}, \hat{\mathbf{w}})$ is estimated by maximum a posteriori according the parameter estimation.

3.2 Likelihood approximation and initialization strategy

Since the observed-data log-likelihood (4) is not tractable, it is approximated by the following harmonic mean [6, chapter 7]:

$$l(\hat{\theta}; \check{\mathbf{x}}) \approx -\ln\left(\frac{1}{Q-B}\sum_{q=B}^{Q} \frac{1}{p(\check{\mathbf{x}}|\hat{\mathbf{x}}^{(q)}, \mathbf{v}^{(q)}, \mathbf{w}^{(q)}; \hat{\theta})}\right) \tag{5}$$

where $(\hat{\mathbf{x}}^{(q)}, \mathbf{v}^{(q)}, \mathbf{w}^{(q)})$ arise independently from $\mathrm{p}(\hat{\mathbf{x}}^{(q)}, \mathbf{v}^{(q)}, \mathbf{w}^{(q)}|\check{\mathbf{x}}; \hat{\theta})$ (they are simulated sequentially as in the SE-Gibbs step with Q iterations after a burn in period of length B), and with:

$$\mathrm{p}(\check{\mathbf{x}}|\hat{\mathbf{x}}^{(q)}, \mathbf{v}^{(q)}, \mathbf{w}^{(q)}; \hat{\theta}) \propto \prod_{ik} \hat{\alpha}_k^{v_{ik}^{(q)}} \prod_{h\ell} \hat{\beta}_\ell^{w_{h\ell}^{(q)}} \prod_{ihk\ell} \mathrm{p}(x_{ih}; \hat{\mu}_{k\ell}, \hat{\pi}_{k\ell})^{v_{ik}^{(q)} w_{h\ell}^{(q)}} \tag{6}$$

where x_{ih} is $\hat{x}_{ih}^{(q)}$ if it corresponds to an unobserved data.

In order to achieve convergence of the SEM-Gibbs algorithm to a global maximum, multiple initialization is used in practice (typically 10 random initializations), and the result maximizing the approximated likelihood will be selected.

3.3 Choice of the number of clusters

In order to select the numbers of clusters, K in rows and L in columns, we propose to adapt to our situation the ICL-BIC criterion developed in [3] in the case of coclustering of categorical data:

$$\text{ICL-BIC}(K, L) = \log \mathrm{p}(\check{\mathbf{x}}, \hat{\mathbf{x}}, \hat{\mathbf{v}}, \hat{\mathbf{w}}; \hat{\theta}) - \frac{K-1}{2} \log n - \frac{L-1}{2} \log d - \frac{KL}{2} \log(nd) \qquad (7)$$

where $\hat{\mathbf{x}}, \hat{\mathbf{v}}, \hat{\mathbf{w}}$ and $\hat{\theta}$ are the respective estimation of the missing data, row partition, column partition and model parameters obtained at the end of the estimation algorithm.

4 Numerical experiments

SEM-Gibbs algorithm validation 30 data sets are simulated according to the following setup: $K = L = 3$ clusters in row and column, $d = 12$ ordinal variables with $m = 5$ levels and $n = 100$ observations. The values of $(\mu_{k\ell}, \pi_{k\ell})$ are given by Table 1.

k/ℓ	1	2	3
1	(1,0.9)	(2,0.9)	(3,0.9)
2	(4,0.9)	(5,0.9)	(1,0.5)
3	(2,0.5)	(3,0.5)	(4,0.5)

Table 1: Parameter values used for experiments.

We use 50 iterations of the SEM-Gibbs algorithm with a burn-in period of 20 iterations. These numbers seem graphically sufficient to achieve stability of the simulations. Figure 1 illustrates the efficiency of the proposed estimation algorithm, by plotting the coclustering results and the following indicators:

- `mu` (resp. `pi`): mean distance between the true $\boldsymbol{\mu}$ (resp. $\boldsymbol{\pi}$) and its estimated value $\hat{\boldsymbol{\mu}}$ (resp. $\hat{\boldsymbol{\pi}}$): $\Delta\boldsymbol{\mu} = \sum_{k=1}^{K} \sum_{\ell=1}^{L} |\mu_{k\ell} - \hat{\mu}_{k\ell}|/(KL)$ (resp. $\Delta\boldsymbol{\pi} = \sum_{k=1}^{K} \sum_{\ell=1}^{L} |\pi_{k\ell} - \hat{\pi}_{k\ell}|/(KL)$),
- `alpha` (resp. `beta`): mean distance between the true α (resp. β) and its estimated value $\hat{\alpha}$ (resp. $\hat{\beta}$): $\Delta\alpha = \sum_{k=1}^{K} |\alpha_k - \hat{\alpha}_k|/K$ (resp. $\Delta\beta = \sum_{\ell=1}^{L} |\beta_\ell - \hat{\beta}_\ell|/L$),
- `ARIr` (resp. `ARIc`): Adjusted Rand Index (ARI) for the row (resp. column) partition.

Ongoing work Further simulation studies (efficiency of the ICL-BIC criterion to select K and L, influence of missing data) and real data analysis will be presented during the conference.

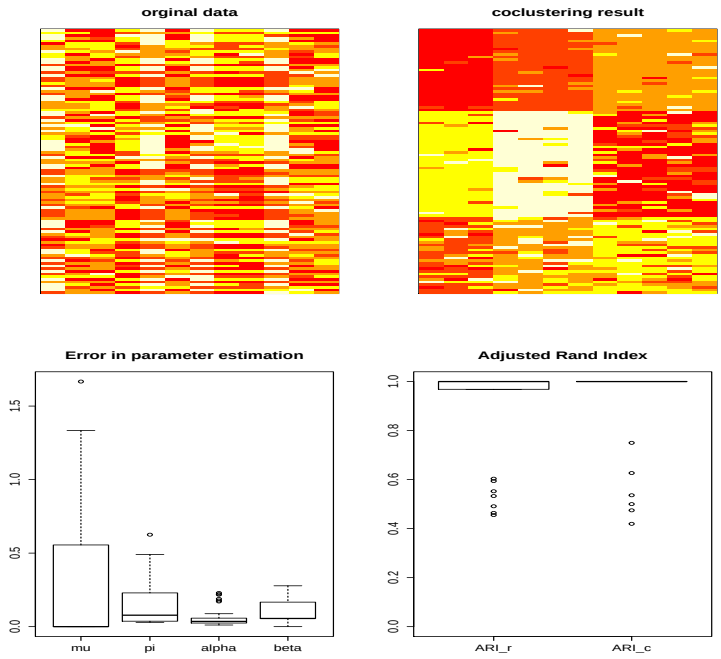

Figure 1: An example of data (top left), coclustering results (top right), error in parameter estimation (bottom left) and ARI for the row and column partitions (bottom right).

References

[1] C. Biernacki and J. Jacques. Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. *Statistics and Computing*, DOI 10.1007/s11222-015-9585-2, 2015.

[2] G. Govaert and M. Nadif. *Co-Clustering*. Wiley-ISTE, 2013.

[3] C. Keribin, V. Brault, G. Celeux, and G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6):1201–1216, 2014.

[4] Govaert G. Keribin C. and Celeux G. Estimation d'un modèle à blocs latents par l'algorithme sem. In *Proceedings of the 42th conference of the French Statistical Society*, Marseille, France, 2010.

[5] E. Matechou, I. Liu, D. Fernandez, M. Farias, and B. Gjelsvik. Biclustering models for ordinal data. Technical report, University of Kent, 2014.

[6] C. P. Robert. *The Bayesian Choice*. Springer, New-York, 2007.